EE 381V: Large Scale Optimization **Fall 2012**

Lecture 24 — November 27

Lecturer: Caramanis & Sanghavi *Scribe: Jahshan Bhatti and Ken Pesyna*

24.1 Mirror Descent

Earlier, we motivated mirror descent as a way to improve the convergence rate of sub-gradient descent with respect to the dimension of the problem. Recall that the bound on the convergence of sub-gradient descent is given by

$$f\left(x_{\text{best}}^{\star}\right) - f^{\star} \leq \frac{L \cdot R}{\sqrt{k+1}},$$

where L is the Lipschitz constant of the function f with respect to $\|\cdot\|_2$ and R is the distance of the initial guess x_0 from the optimal point $x^{\star}$: $\|x_0 - x^{\star}\|_2$. Also, recall that the sub-gradient update is given by

$$\begin{aligned} x^{+} &= \operatorname{Proj}_{\mathscr{X}}\left(x - \gamma_t g\right) \\ &= \arg\min_{u \in \mathscr{X}} \left[\left\langle \gamma g - \nabla\omega\left(x\right), u\right\rangle + \omega\left(u\right)\right], \end{aligned}$$

where $g \in \partial f\left(x\right)$ and $\omega\left(u\right) = \frac{1}{2}\|u\|_2^2$ is the "distance generating function" (DGF) that is continuous, differentiable, and strongly convex with respect to $\|\cdot\|_2$. The main idea of mirror descent is to replace $\omega(u) = \frac{1}{2}\|u\|_2$ with some other DGF so that the bounds are replaced by $L \to L^f$ and $R \to R^f$ where R^f is the "size of set" as measured by the new Bregman divergence DGF $\omega\left(\cdot\right)$.

$$f\left(x_{\text{best}}^{\star}\right) - f^{\star} \leq \frac{L^f \cdot R^f}{\sqrt{k+1}},$$

Note that $\omega\left(\cdot\right)$ should be α-strongly convex with respect to the norm $\|\cdot\|$ used.

24.1.1 Analysis of Convergence

In order to analyze the convergence of mirror descent, we'll first consider the Lyapunov function $(\|x_k - x^{\star}\|_2)$ used in the Euclidean case for sub-gradient descent. Then we'll discuss how the equation will change for mirror descent.

The key inequality for the convergence analysis of sub-gradient descent is a guaranteed decrease in the Lyapunov function, which is given, for any $u \in \mathscr{X}$, by

$$\frac{1}{2}\|x - u\|_2^2 - \frac{1}{2}\|x_+ - u\|_2^2 \geq \gamma\left\langle g, x - u\right\rangle - \frac{1}{2}\gamma^2\|g\|_2^2. \tag{24.1}$$

Recall from last lecture that the Bregman divergence of is given by

$$D(u,v) = \omega(u) - \omega(v) - \langle \nabla\omega(v), u - v\rangle .$$

Therefore, the analog to the Lyapunov key inequality in (24.1) for mirror descent, can be formed by replacing $\|u-v\|_2^2$ by the $D(u,v)$ and can be reformulated, for some iteration t, as

$$D(u,x_t) - D(u,x_{t+1}) \geq \gamma_t \langle g_t, x_t - u\rangle - \frac{1}{2\alpha}\gamma_t^2 \|g_t\|_\star^2, \tag{24.2}$$

(For $w(u) = \frac{1}{2}\|u\|_2^2$, 24.2 this is exactly what we had in 24.1.)

Eq. 24.2 can be rewritten as

$$\underbrace{[\langle \nabla\omega(x_t), x_t - u\rangle - \omega(x_t)]}_{H_u(x_t)} - \underbrace{[\langle \nabla\omega(x_{t+1}), x_{t+1} - u\rangle - \omega(x_{t+1})]}_{H_u(x_{t+1})}$$
$$\geq \gamma_t \langle g_t, x_t - u\rangle - \frac{1}{2\alpha}\sum \gamma_t^2 \|g_t\|_\star^2 . \tag{24.3}$$

To complete the convergence analysis, recall, for any $u \in \mathscr{X}$ and for some iteration t,

$$\begin{aligned} f(u) &\geq f(x_t) + \langle g_t, u - x_t\rangle \\ \gamma_t(f(x_t) - f(u)) &\leq \gamma_t \langle g_t, x_t - u\rangle . \end{aligned}$$

Then, summing (24.3) from $t=0$ to $t=T$ forms a telescoping sum that yields

$$\begin{aligned} \sum_{t=0}^{T} \gamma_t \langle g_t, x_t - u\rangle &\leq \underbrace{H_u(x_0) - H_u(x_T)}_{\Theta} + \frac{1}{2\alpha}\sum \gamma_t^2 \|g_t\|_\star^2 \\ \sum \gamma_t \underbrace{(f(x_t) - f(u))}_{f\left(x_{\text{best}}^T\right) \leq f(x_t)} &\leq \\ \underbrace{\left(f\left(x_{\text{best}}^T\right) - f(u)\right)}_{\text{Let } u = x^\star} \sum \gamma_t &\leq \Theta + \frac{1}{2\alpha}\sum \gamma_t^2 \|g_t\|_\star^2 \\ f\left(x_{\text{best}}^T\right) - f^\star &\leq \frac{\Theta + \frac{1}{2\alpha}\sum \gamma_t^2 \|g_t\|_\star^2}{\sum \gamma_t}, \end{aligned}$$

where Θ is the upper bound on $\|x^\star - x_0\|_2^2 = \operatorname{diam}\mathscr{X}$, or generally the "size of $\mathscr{X}$ measured by $D(\cdot,\cdot)$," the Bregman divergence. $f(x_{\text{best}}^T)$ represents the closest that $f(\cdot)$ gets to $f^\star$ over the entire time interval.

Now, consider a specific value for the step size γ_t at each iteration t,

$$\gamma_t = \frac{\sqrt{\Theta \cdot \alpha}}{\|g_t\|_\star \cdot \sqrt{t}}.$$

Given this step size, it is possible to show that the error in the optimal function value $\epsilon_T \triangleq f\left(x^T_{\text{best}}\right) - f^\star$ can be bounded as follows:

$$\epsilon_T \leq O\left(1\right) \frac{\sqrt{\Theta} L^f_{\|\cdot\|}}{\sqrt{2}\sqrt{T}}.$$

Note that when using the l2-norm $L_{\|\cdot\|} = L_{\|\cdot\|_2}$ and for the DGF $w(\cdot) = \frac{1}{2}\left\|\cdot\right\|_2$, we recover an upper bound that is exactly what we had for subgradient descent.

24.1.2 Simplex Mirror Descent versus Subgradient Descent

A natural question to ask is: when does the convergence rate of mirror descent exceed the convergence rate of standard subgradient descent. Here's one scenario: Let the feasible set be the simplex set scaled by k: $X \in \Delta^+_n(k)$, let the distance generating function be$\omega\left(x\right) = \sum x_i \ln\left(x_i\right)$ and let our norm be the l1-norm , $\left\|\cdot\right\| = \left\|\cdot\right\|_1$. For these parameters, the mirror descent update is easy. The modulus of strong convexity with respect to the l1-norm is $\alpha = O(1)/R^2$ and the upper bound $\Theta \leq O\left(1\right)\ln\left(n\right)$. Thus, for this scenario, the upper bound on the convergence rate is

$$\epsilon_T \leq O\left(\sqrt{\ln\left(n\right)}\right) \frac{L^F_{\|\cdot\|_1} R}{\sqrt{T}}.$$

Now that we have a convergence bound for Mirror Descent with a simplex set, we can compare this to the bound for subgradient descent, i.e. using the Euclidean norm. We consider the following ratio comparing the convergence error of the two methods:

$$\frac{\epsilon_{MD-Simplex}}{\epsilon_{SD}} = \frac{O\left(\sqrt{\ln\left(n\right)}\right) \frac{L^F_{\|\cdot\|_1} R}{\sqrt{T}}.}{\frac{L_{\|\cdot\|_2} \cdot R}{\sqrt{T}}} \tag{24.4}$$

$$= \underbrace{\frac{O\left(\sqrt{\ln\left(n\right)}\right)}{1}}_{(I)} \cdot \underbrace{\frac{\max_X \left\|x-y\right\|_1}{\max_X \left\|x-y\right\|_2}}_{(II)} \cdot \underbrace{\frac{L^f_{\|\cdot\|_1}}{L^f_{\|\cdot\|_2}}}_{(III)} \tag{24.5}$$

We can break apart Eq. 24.4 into three terms so that we can classify them as either favoring Mirror Descent or Subgradient descent.

- (I) Always Favors subgradient descent (i.e. the Euclidean norm) since the numerator is ≥ 1
- (II) This term represents the error in the initial guess. This always favors subgradient descent since the l1-norm-based numerator is at worst equal to $\sqrt{n}$, where n is the

dimension of the system and at best is 1. The denominator is always 1 since the set for subgradient descent is the Euclidean ball. Consequently this ratio is always between 1 and $\sqrt{n}$:$1 \leq ratio \leq \sqrt{n}$

- (III) Favors Mirror Descent with a simplex set over subgradient descent with a Euclidean set ($\frac{1}{\sqrt{n}} \leq ratio \leq 1$)

From this analysis we can make the following conclusions:

1. If our set $\mathscr{X}$ is a Euclidean ball and our function f is sensitive to $O\left(1\right)$ coordinate and subgradient descent much better by a factor of $\sqrt{n\ln\left(n\right)}$

2. If our set$\mathscr{X}$ is a simplex and our function f is sensitive to $O\left(n\right)$ coordinates, then Mirror Descent is better by a factor of: $\frac{\sqrt{n}}{\sqrt{\ln(n)}}$

24.2 Algorithms that use the Dual

Recall the concept of duality:

The primal of the problem is

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{subject to} \quad & h(x) \leq 0 \\ & Ax = b \end{aligned}$$

The Lagrangian for the problem is:

$$\mathscr{L}_{\lambda \geq 0}(x, \lambda, \nu) \;\; = \;\; f(x) + \lambda^T h(x) + \nu(Ax - b)$$

The dual objective of the problem is

$$g(\lambda, \nu) \;\; = \;\; \min_{x} \mathscr{L}(x, \lambda, \nu)$$

The solution to the dual is

$$\lambda^\star, \nu^\star = arg \max_{\lambda \geq 0, \nu} g(\lambda, \nu)$$

from which we can recover the optimal solution to the primal by:

$$x^\star = arg \min_{x} \mathscr{L}(x, \lambda^\star, \nu^\star)$$

24.2.1 Primal and Dual Decomposition

Oftentimes, it becomes possible to exploit the problem structure to parallelize the solution for faster processing. However during parallelizing we often have to deal with one of two complications:

- Coupled variables
- Coupled constraints

24.2.1.1 Primal Decomposition

The primal decomposition master problem given a coupled variable can be posed as follows:

$$\min_{y} \quad \phi_1(y) + \phi_2(y)$$

where ϕ_1 and ϕ_2 are the subproblems defined as follows:

$$\begin{aligned} \phi_1(\mathrm{y}) &= \min\nolimits_{x_1} \quad f_1(x,y) \\ \phi_2(y) &= \min\nolimits_{x_2} \quad f_2(x,y) \end{aligned}$$

The master problem can then be solved by iterating between the master problem and the individual subproblems. The master problem is always feasible, meaning that the constraints are always met during the minimization process

24.2.1.2 Dual Decomposition

Similar to primal decomposition, it is possible to parallelize the dual minimization problem by breaking it into subproblems. For following primal master problem,

$$\begin{aligned} \min_{x_1 y_1 x_2 y_2} \quad & f_1(x_1,y_1) + f_2(x_2,y_2) \\ \text{subject to} \quad & y_1 = y_2 \end{aligned}$$

the Lagrangian can can be formed as follows

$$\mathscr{L}(x_1,y_1,x_2,y_2) \quad = \quad f_1(x_1,y_1) + f_2(x_2,y_2) + \lambda(y_1 - y_2)$$

and then split into two subproblems as follows

$$\begin{aligned} &\text{subproblem 1} \quad \min\nolimits_{x_1 y_1} \quad f_1(x_1,y_1) + \lambda y_1 \\ &\text{subproblem 2} \quad \min\nolimits_{x_2 y_2} \quad f_2(x_2,y_2) - \lambda y_2 \end{aligned}$$

The Lagrangian multiplier is used to enforce the primal constraint and can be updated using the following update equation

$$\lambda_+ \quad = \quad \lambda - \alpha(y_2 - y_1)$$

where α is the step size. This update equation is effectively gradient ascent on the dual w.r.t λ. The dual can then be solved by iterating between the subproblems, the update equation. Unlike primal decomposition, dual decomposition can be infeasible at times during the minimization process, as the coupled variable can be different in each subproblem. This concludes coupled variables. Next lecture will review coupled variables and introduce coupled constraints.